bio-protocol Preprint

Not Peer Reviewed
Ask a question online

ATAC-seq and ChIP-seq data analysis on mouse strains data

Zeyang Shen Marten A Hoeksema Christopher K. Glass

Updated date: Jul 8, 2021

An abbreviated version of this protocol was published in Science Advances in Jun 2021

Mechanisms underlying divergent responses of genetically distinct macrophages to IL-4

DOI: 10.1126/sciadv.abf9808

Detailed protocol

Abstract

Assay for Transposase Accessible Chromatin using sequencing (ATAC-seq) analysis is used to investigate open chromatin landscapes in various cell types and disease states. In combination with Chromatin ImmunoPrecipitation sequencing (ChIP-seq), one can further investigate the chromatin landscape and define active or repressed gene enhancers and promoters. ChIP-seq is also used to define transcription factor binding patterns throughout the genome. In this study, we have used macrophages from genetically distinct mouse strains to study the effects of genetic variants on transcription factor binding and enhancer activity. In this protocol, we describe how to analyze ATAC-seq and ChIP-seq from samples of various mouse genomes.

Data availability

All the sequencing data generated from this study (Hoeksema et al., 2021) and used for the analysis described below is publicly available in the GEO database: GSE159630. These data can be visualized through the UCSC browser under session: https://genome.ucsc.edu/s/zeyang/IL4_mm10_strains.

Required software

- Bowtie2 v2.2.9: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
- MMARGE v1.0: https://github.com/vlink/marge
- HOMER v4.11.1: http://homer.ucsd.edu/homer/data/software/
- IDR v2.0.3: https://github.com/nboley/idr
- MAGGIE v1.1.1: https://github.com/zeyang-shen/maggie

Data analysis

A. Custom genome generation for mouse strains

Custom genomes were generated for BALB/cJ, NOD/ShiLtJ, PWK/PhJ, and SPRET/EiJ mice from the C57BL/6J or mm10 genome as before (Link et al., 2018a) using MMARGE (Link et al., 2018b) and the VCF files from the Mouse Genomes Project (Keane et al., 2011). Specific instructions can be found in MMARGE documentation (https://github.com/vlink/marge/blob/master/MMARGE_documentation.pdf) Chapter 8 Getting started with data processing.

B. Sequencing data processing

1. Mapping data. ATAC-seq and ChIP-seq data were mapped to the genome of corresponding mouse strain using bowtie2 with default parameters (Langmead and Salzberg, 2012).

```
bowtie2 [fastq file] -p 8 -x [mouse strain genome bowtie index prefix] > [pre-shifting sam file] 2> [log file]
```

2. Shifting to the mm10 genome. The mapped data was then shifted to the mm10 genome using the MMARGE 'shift' function (Link et al., 2018b) for downstream comparative analyses across different mouse strains.

```
MMARGE.pl shift -sam -dir . -files [pre-shifting sam file] -data_dir [directory storing mouse strain genomes] -ind [mouse strain genome name]
```

3. Creating HOMER tag directory. A HOMER tag directory was created from each shifted SAM file as the prerequisite for using HOMER (Heinz et al., 2010) for downstream analysis.

```
makeTagDirectory [tag directory name] -genome mm10 -checkGC [post-shifting sam file]
```

C. Peak calling

We called peaks for ATAC-seq data and transcription factor ChIP-seq data to identify open chromatin regions and transcription factor binding sites, respectively.

1. Calling unfiltered peaks with HOMER. Based on the HOMER tag directories created from mapped sequencing data, we first used HOMER to call unfiltered 200-bp peaks based on each replicate:

```
findPeaks [tag directory] -style factor -L 0 -C 0 -fdr 0.9 -o [unfiltered peak file name] -i [tag directory of input sample] -size 200
```

Input tag directory *-i* can be ignored when dealing with ATAC-seq data.

2. Finding reproducible peaks with IDR. Since our data include two replicates for every condition of every mouse strain, we then ran IDR (Li et al., 2011) on replicates of the same sample using an IDR threshold equal to 0.05. Before running IDR, HOMER peak files (http://homer.ucsd.edu/homer/ngs/peaks.html) need to be converted into BED format first (https://genome.ucsc.edu/FAQ/FAQformat.html#format1) mainly by moving the peak IDs from the 1st column to the 4th column.

```
idr --samples [unfilteredPeakFileForReplicate1 unfilteredPeakFileForReplicate2] --output-file [output folder path] --plot --idr-threshold 0.05
```

D. Cross-sample comparison

1. Merging peaks. For comparisons across multiple samples (e.g., different time points, mouse strains, transcription factors), we first merged the specified set of peaks using HOMER mergePeaks:

```
mergePeaks -d given [IDRpeakFile1 IDRpeakFile2 ...] > [merged peak file name]
```

2. Quantifying signal intensities.

- Levels of histone modifications and RNA polymerase II: quantified within +/- 500 bp around the centers of ATAC-seq reproducible peaks using HOMER annotatePeaks.pl script:

  ```
  annotatePeaks.pl [merged peak file] mm10 -norm 1e7 -d [tagDirectory1 tagDirectory2 ...] -size -500,500 > [annotated peak file]
  ```

- Transcription factor binding intensity: quantified within +/- 150 bp around the identified ChIP-seq peaks using HOMER annotatePeaks.pl script:

  ```
  annotatePeaks.pl [merged peak file] mm10 -norm 1e7 -d [tagDirectory1 tagDirectory2 ...] -size -150,150 > [annotated peak file]
  ```

3. Visualizing average profiles. To visualize the average profiles of multiple datasets around a certain set of peaks, we used HOMER annotatePeaks.pl to help compute the histograms of 20-bp bins within +/- 2000 bp regions:

```
annotatePeaks.pl [merged peak file] mm10 -norm 1e7 -size 4000 -hist 20 -ghist -d [tagDirectory1 tagDirectory2 ...] > [histogram file]
```

E. Motif enrichment analysis

Given a certain set of peaks, we used HOMER findMotifsGenome.pl to identify de novo motifs and the matched known motifs. The background sequences were either default random sequences or another set of peaks from a comparative condition.

```
findMotifsGenome.pl [peak file] mm10 [output folder path] -size 200 -mask -p 8 -bg [comparative peak file]
```

To use random backgrounds, the parameter *-bg* can be ignored.

F. Motif mutation analysis

The detailed instructions can be found at https://github.com/zeyang-shen/maggie/wiki/Identify-functional-motifs-from-ChIP,-ATAC,-DNase-seq-peaks In brief, based on the peaks identified from each mouse strain, we first found differential peaks between strains through pairwise comparisons (Step 2). Then we extracted sequences of these differential peaks from respective mouse strain genomes (Step 3) and ran MAGGIE (Step 4; Shen et al, 2020).

References

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell, 38(4), 576–589.

Hoeksema, M.A., Shen, Z., Holtman, I.R., Zheng, A., Spann, N.J., Cobo, I., Gymrek, M., and Glass, C.K. (2021). Mechanisms underlying divergent responses of genetically distinct macrophages to IL-4. Sci Adv 7.

Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477, 289-294.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357-359.

Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. The annals of applied statistics 5, 1752-1779.

Link, V.M., Duttke, S.H., Chun, H.B., Holtman, I.R., Westin, E., Hoeksema, M.A., Abe, Y., Skola, D., Romanoski, C.E., Tao, J., et al. (2018a). Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide Mechanisms that Control Transcription Factor Binding and Function. Cell 173, 1796-1809 e1717.

Link, V.M., Romanoski, C.E., Metzler, D., and Glass, C.K. (2018b). MMARGE: Motif Mutation Analysis for Regulatory Genomic Elements. Nucleic Acids Res 46, 7006-7021.

Shen, Z., Hoeksema, M.A., Ouyang, Z., Benner, C., Glass, C. K. (2020). MAGGIE: leveraging genetic variation to identify DNA sequence motifs mediating transcription factor binding and function. Bioinformatics 36, i24-i32.

How to cite:(Readers should cite both the Bio-protocol preprint and the original research article where this protocol was used)

1. Shen, Z. , Hoeksema, M. A. and Glass, C. (2021). ATAC-seq and ChIP-seq data analysis on mouse strains data. Bio-protocol Preprint. bio-protocol.org/prep1272.

2. Hoeksema, M. A., Shen, Z., Holtman, I. R., Zheng, A., Spann, N. J., Cobo, I., Gymrek, M. and Glass, C. K.(2021). Mechanisms underlying divergent responses of genetically distinct macrophages to IL-4. Science Advances 7(25). DOI: 10.1126/sciadv.abf9808